图书情报工作
第66卷 第6期 2022年3月

基于信息增益与相似度的专利关键词抽取算法评价模型*

■ 俞琰[1,2]　鞠鹏[1]　尚明杰[1]

[1] 南京工业大学信息管理与技术研究所　南京 210009　[2] 东南大学成贤学院计算机工程系　南京 211816

摘　要：[目的/意义]针对目前专利关键词抽取算法评价中主要采用抽取的关键词与专家人工标注关键词进行匹配存在的问题，提出一种基于信息增益与相似度的专利关键词抽取算法评价模型。[方法/过程]提出的评价模型从内部和外部两个层面评估专利关键词抽取算法的准确性。其中，内部评价模型度量待评价算法抽取的每个关键词的信息增益，以评估被抽取的关键词的新颖性与创造性；外部评价模型使用待评价算法抽取的关键词集表示专利，计算相关专利的相似度，衡量算法抽取的关键词描述专利主题的有效性。[结果/结论]通过评价模型有效性验证实验与评价模型应用实证研究，结果表明提出的基于信息增益与相似度的评价模型具有可行性与有效性。

关键词：专利　关键词抽取　评价　信息增益　相似度
分类号：G202
DOI：10.13266/j.issn.0252-3116.2022.06.012

1　引言

专利关键词是表明专利文献主题内容的一组词或者短语，被广泛应用于专利分类[1]、新兴技术监测[2]、专利检索[3]、专利聚类[4]等专利分析之中。而专利通常不包含关键词，需要人工标引。由于专利文献篇幅较长、内容专业，且近年来数量急剧增长，使得人工标引专利关键词方法已无法满足专利文献分析的需要。因此，利用计算机自动、高效、准确地抽取专利关键词是一个重要的研究课题。目前专利关键词抽取研究主要集中于专利关键词抽取算法的改进，而研究中各种改进算法的评价通常将算法抽取的关键词与专家人工标注的关键词进行匹配，以评估抽取算法的有效性。然而，依靠专家人工标注专利关键词费时费力、标注数量有限，存在领域局限性、语言依赖性、主观性等问题，使得专利关键词抽取算法无法有效地被评价，阻碍专利关键词抽取算法的进一步深入研究。

据此，本文提出一种基于信息增益与相似度的专利关键词抽取算法评价模型，以缓解目前评价方法的不足。提出的评价模型从内部和外部两个层面，分别评价算法抽取关键词的准确性。其中，内部评价通过计算带评价算法抽取的每个关键词的信息增益，度量被抽取关键词的新颖性和创造性；外部评价使用算法抽取的关键词集表示专利文献，计算相关专利的相似度，以评估抽取的专利关键词表示专利主题内容的准确性。此外，本文提出的评价模型不仅适用于专利文献关键词抽取算法评价，也适用于学术文献等相关文献关键词抽取算法评价。

总的，本文贡献如下：

（1）鉴于目前专利关键词抽取算法评价方法存在的不足，提出基于信息增益与相似度的专利关键词抽取算法评价模型（第3节）。

（2）进行评价模型有效性实验，结果表明本文提出的评价模型的有效性（第4节）。

（3）利用本文提出的评价模型进行应用实证分析，评价3种专利关键词抽取策略，应用结果表明本文提出的评价模型的有效性与可行性（第5节）。

2　相关研究

关键词抽取算法评价主要考察算法抽取的关键词集合反应文献主题内容的准确程度。评价方法可分为内部评价方法（intrinsic evaluation）和外部评价方法（extrinsic evaluation）两大类[5]。

内部评价方法将算法抽取的关键词与正确关键词进行匹配，判断被抽取的关键词是否正确，然后使用评分指标评价抽取算法的准确性。

* 本文系国家社会科学基金项目“大数据时代支持创新设计的多维度多层次专利文本挖掘研究”（项目编号：17BTQ059）研究成果之一。
作者简介：俞琰，教授，博士，E-mail：yuyanyuyan2004@126.com；鞠鹏，硕士研究生；尚明杰，硕士研究生。
收稿日期：2021-07-08　修回日期：2021-10-30　本文起止页码：108-117　本文责任编辑：徐健

俞琰，鞠鹏，尚明杰. 基于信息增益与相似度的专利关键词抽取算法评价模型[J]. 图书情报工作,2022,66(6):108－117.

其中,匹配通常采用精确匹配法,即将算法抽取的关键词与文献作者或者专家人工标注的关键词(称为金标准关键词)进行比较。然而,精确匹配比较条件过于严格,忽略了语义关联,如两个关键词之间为同义词或者部分匹配等情况,造成评价结果的不可靠。为此,一些研究在精确匹配的基础上,添加模糊匹配方法作为补充。如,采用编辑距离计算关键词间的词形相似度[6],采用概率模型计算关键词间的语义相似度[7],以及综合考虑词形与语义信息的相似度进行匹配[8]。

内部评价方法中使用最广泛的评分指标为查准率(precision, P)、查全率(recall, R)和F1值(F1-score)[9－11]。查准率度量匹配关键词占所抽取关键词比率;查全率衡量匹配关键词占金标准关键词比率;F1值为查准率和查全率的加权平均。然而,这些评分指标没有考虑被抽取关键词的顺序。实际上,如果匹配关键词具有更高的排名,则该抽取算法具有更高的准确性。为此,一些研究将评分指标进行改进,根据被抽取关键词排名顺序进行评分,如 Precision@K[12]指标考虑前K个被抽取的关键词,计算被抽取关键词的查准率,K通常取1、3、5、10等值;平均倒数排名(mean reciprocal rank, MRR)[13－15]度量第1个匹配的关键词的排名情况;二元偏好度量(Binary preference measure, Bpref)[16－17]计算提取结果中错误提取的词语的排名情况。

外部评价方法将算法抽取的关键词用于一个特定的应用之中。通过度量这些应用的性能,间接地评价算法抽取关键词的效果。如文本分类[18]、聚类或检索[19],根据这些特定任务的结果来评价关键词抽取方法的效果。由于外部评价方法是针对特定任务,所以任务中所使用语料的质量、规模、任务采用的算法均对评价结果有很大影响,且特定任务本身的计算量通常会超过关键词抽取算法本身的计算量,从而使得评价速度难以满足实际需要。因此,目前外部评价方法较少用于关键词抽取算法效果评价。

目前内部评价方法是使用最为广泛的评价方法。然而,该评价方法具有较大局限性。其中,一些文献,尤其是专利文献本身没有关键词,需要在评价抽取算法时,人工标注关键词作为金标准关键词。而人工标注关键词工作量大、具有主观性与任意性。语料类型、标注粒度、标注人员专业素质等条件不同,在不同数据集上的标注结果存在较大差异。因此需要多位领域专家参与标注,并计算专家之间标注关键词的一致程度,如通常使用Kappa统计衡量不同标注的一致性。虽然目前有一些开源的标注数据集[20－21],但这些数据集标注可靠性较差[5],通常为特定领域的英文数据集,且没有专门的专利关键词标注数据集。

因此,研究者探索在没有金标准关键词的情况下进行关键词抽取算法评价。如章成志[8]使用词语的频度和位置信息抽取文献关键词作为金标准关键词,然后与待评价算法抽取的关键词集合进行相似度比较,以评价算法准确性。然而,该评价方法基于这样一个假设,即,使用词语频度和位置信息抽取的文献关键词为金标准,实际上,仅使用词语频度和位置信息得到的关键词并不能准确反应文献内容。反之,若使用词语频度和位置信息得到的关键词能够作为金标准关键词,准确反应文献主题内容,则不需要关键词抽取算法的改进研究。

基于以上分析,本文提出一种新的评价模型以缓解目前评价方法存在的不足,促进专利关键词抽取算法研究的进一步发展。

3 评价模型

3.1 评价模型原理

评价模型从内部评价和外部评价两个层面,分别提出基于信息增益的内部评价模型和基于相似度的外部评价模型。内部评价使用信息增益分别度量每个被抽取关键词的有效性;外部评价则将算法抽取的关键词集作为一个整体,表示专利文献主题内容,计算专利间的相似度,以衡量关键词集刻画文献主题内容的准确性。

3.1.1 基于信息增益的内部评价模型原理

专利关键词应表明专利主题内容,体现专利的新颖性和创造性。因此专利关键词应携带尽可能多的区分已有专利的信息量。信息增益(information gain, IG)表示在某一条件下随机事件不确定性的减少程度,从而表明该条件所携带的信息量。当被抽取的关键词使得已有系统的不确定性减少越多,则该被抽取的关键词包含的信息量越大。算法的信息增益则可用该算法抽取的关键词的信息增益均值表示,以度量算法抽取关键词新颖性和创造性的能力。

图1为基于信息增益的内部评价模型原理示例。在图1中,待评价算法1和算法2分别从目标专利抽取2个关键词,通过构建相关专利数据集,计算算法1抽取的关键词“锂离子”和“三元”的信息增益分别为0.41和1.85,因此,算法1的信息增益为1.13;算法2抽取的关键词“发明”和“电池”的信息增益分别为

0.01和0.11,因此算法 2 的信息增益为0.06。表明使用算法 1 抽取的关键词携带更多的信息量,更能体现目标专利的新颖性与创造性。

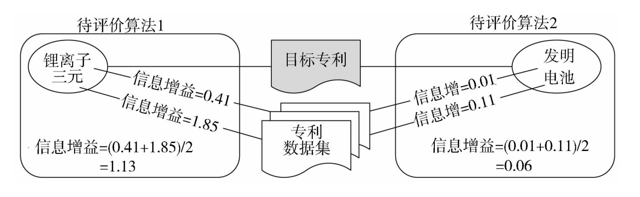


图 1 基于信息增益的内部评价模型原理示例

根据上述分析可知,基于信息增益的内部评价模型与传统的内部评价模型的异同点。相同之处在于,两者均对算法抽取的每个关键词进行评价。不同之处在于,传统的内部评价模型对每个抽取的关键词与金标准关键词进行匹配,而本文提出的基于信息增益的内部评价模型则无需人工标注关键词,通过度量每个被抽取关键词的信息增益,衡量算法的有效性。

3.1.2 基于相似度的外部评价模型原理

专利关键词反应专利主题内容,因此外部评价模型使用抽取的关键词表示专利,通过关键词集计算专利间的相似度,以检测其刻画专利主题内容的准确性。专利中的创新不是孤立事件,创新在一定程度上体现技术的承接。专利审查员会对照相似专利,标注相关相似专利为引用专利,故引用专利在一定程度上刻画目标专利与引用专利的相似性[22-24]。研究表明,目标专利与引用专利比目标专利与随机专利具有更高的相似度[22-24]。因此,本文利用算法抽取的关键词集计算目标专利与引用专利的相似度、目标专利与随机专利的相似度,以检测算法刻画专利提取专利主题的准确性。

图 2 为基于相似度的外部评价模型原理示例。待评价算法 1 和待评价算法 2 分别从目标专利、引用专利和随机专利抽取 3 个关键词。待评价算法 1 抽取的关键词集中,目标专利与引用专利具有相同的关键词“锂离子”“正极”,其相似度为 2,目标专利与随机专利仅有一个共同关键词“锂离子”,其相似度为 1;待评价算法 2 中,目标专利与引用专利的相同关键词为“发明”和“电池”,相似度为 2,目标专利与随机专利的相同关键词为“发明”和“电池”,相似度为 2。算法 1 得到的目标专利与引用专利的相似度大于目标专利与引用专利的相似度,符合直觉,更准确地刻画了专利主题内容,因此算法 1 比算法 2 具有更好的关键词抽取效果。

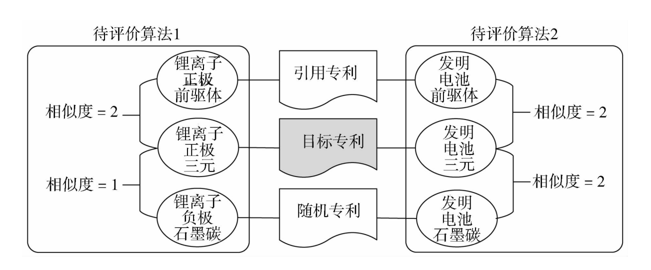


图 2 基于相似度的外部评价模型原理示例

根据上述内容分析基于相似度的外部评价模型与传统的外部评价模型的异同点。相同之处在于,两者均使用算法抽取的关键词表示文献主题,应用于具体应用中,以评价算法的准确性。不同之处在于,相对于传统的外部评价模型,本文提出的基于相似度的外部评价模型不需要复杂的外部应用,通过计算目标专利与引用专利,以及目标专利与随机专利的相似度,评价算法的准确性,比传统方法更具实用性。

3.2 评价模型具体描述

3.2.1 基于信息增益的内部评价方法

信息增益表示某一条件 y 下,随机事件 X 不确定性的减少程度,从而表明该条件 y 所携带的信息量:

$$IG(y) = H(X) - H(X|y) \qquad 式(1)$$

其中,H 表示信息熵,$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$,$p(x_i)$代表随机事件 X 为 x_i 的概率,$H(X)$用于衡量系统的不确定性,系统不确定性越大,则信息熵越高;$H(X|y) = -\sum_{i=1}^{n} p(x_i|y) \log p(x_i|y)$,$H(X|y)$表示条件 y 下,系统的不确定性。信息增益通过衡量条件 y 出现前后信息熵的变化,表明其通过引入信息的多寡以消除系统不确定的程度。当新的信息被接受和处理,系统不确定性减少。如果信息消除的不确定性越大,则其包含的信息量也就越大。

据此,本文构建如图 3 所示的专利数据集 $S = \{C_1, C_2, \cdots, C_n\}$,$C_i(i = 1, 2, \cdots, n)$表示某专利类别,$C_i = \{d_{i,1}, d_{i,2}, \cdots, d_{i,m}\}$,$d_{i,j}(j = 1, 2, \cdots, m)$为 C_i 类别中的一个专利,专利关键词应携带尽可能多的信息,以区分不同专利类别,以及相同专利类别中不同专利,消除尽可能多的不确定性。进一步地,本文将消除专利类别间不确定的信息增益称为类别信息增益(Information Gain of Class, IGC),把消除同一类别内不同专利文献的不确定的信息增益称为文档信息增益(information gain of document, IGD)。

具体地,给定目标专利 $d_{i,j}$ 抽取的关键词 w,类别信息增益度量 w 携带的信息量,消除专利类别间不确定性的能力:

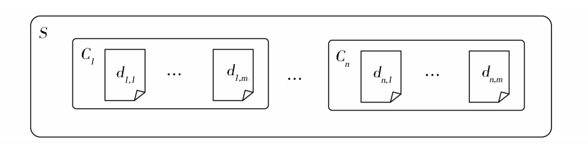


图3　基于信息增益的专利数据集构建

$$IGC(w)=(-\sum_{i=1}^{n}p(C_i)\log p(C_i))-(-\sum_{i=1}^{n}p(C_i|w)\log p(C_i|w))\quad 式(2)$$

其中，$p(C_i)$表示类别 C_i 出现的概率，$p(C_i|w)$表示词语 w 出现后，类别 C_i 出现的概率。正确的专利关键词应携带较多的信息，尽可能多地消除专利类别间的不确定性。

文档信息增益度量 w 消除与目标专利 $d_{i,j}$在同一类别 C_i 的不同专利文献的不确定性的能力：

$$IGD(w)=(-\sum_{j=1}^{m}p(d_{i,j})\log p(d_{i,j}))-(-\sum_{j=1}^{m}p(d_{i,j}|w)\log p(d_{i,j}|w))\quad 式(3)$$

其中，$p(d_{i,j})$表示专利 $d_{i,j}$出现概率，$p(d_{i,j}|w)$表示词语 w 出现后，专利 $d_{i,j}$出现的概率。关键词应携带较多的信息，消除同一类别内专利文献间的不确定性。

最终，专利关键词尽可能消除类别间以及类别内的不确定性，因此，使用类别信息增益和文档信息增益之积表示词语的信息增益 $IG(w)$：

$$IG(w)=(IGC(w)+\alpha)\times(IGD(w)+\alpha)\quad 式(4)$$

其中，α 为很小的实数，避免 IG 为0。

一个词语的信息增益越大，该词语所携带的信息越多，更有可能是专利的关键词。图4为专利关键词内部评价方法示例，每篇专利中包含若干词语。目标专利为 $d_{1,1}$，为一个关于锂离子电池正极材料相关发明专利。表1为对应的词语“发明”“电池”和“三元”的信息增益值。由表1可见，词“发明”由于在各类别以及目标类别中均有大量均匀出现，其类别信息增益和文档信息增益均很小；词“电池”在目标类别的各专利文档中大量均匀出现，而在其他类别中较少出现，具有较低的文档信息增益值和较高的类别信息增益值；词“三元”仅在目标类别的少数专利文档中出现，具有较高的文档信息增益值和类别信息增益值，使得其最终的信息增益值最高，从而具有较强的区分类别间和类别内专利的能力。由此可见，本文提出的信息增益能够较好地刻画一个词给已有专利文献所带来的信息增量，具有一定的合理性。

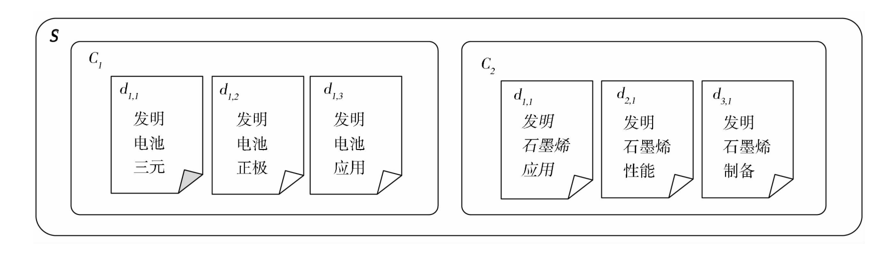


图4　基于信息增益的内部评价方法示例

表1　信息增益计算示例

词	类别信息增益	文档信息增益	信息增益
发明	1.0－1.0 ＝0	1.58－1.58 ＝0	0.01
电池	1.0－0 ＝1.0	1.58－1.58 ＝0	0.11
三元	1.0－0 ＝1.0	1.58－0 ＝1.58	1.85

注：(式(4)中a ＝0.1)

3.2.2　基于相似度的外部评价方法

基于相似度的外部评价方法数据集构建如图5所示。给定目标专利，使用目标专利和引用专利构造相似专利对，使用目标专利和随机专利构造随机专利对，使用抽取的关键词集分别计算相似专利对和随机专利对的相似度。若抽取的关键词能够准确表示专利主题内容，则相似专利对的相似度应大于随机专利相似对的相似度。

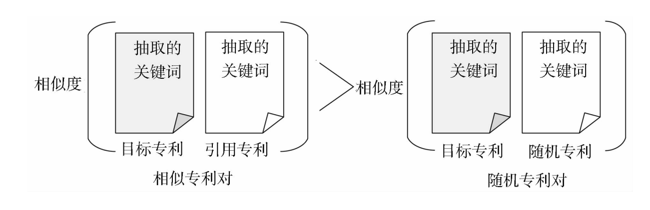


图5　相似专利对与随机专利对构建

具体地，给定目标专利 d，与其引用专利 d_c，以及随机专利 d_r，采用抽取的关键词表示专利文档，计算两个专利 d_i 和 d_j 的相似度 sim：

$$sim(d_i,d_j)=|d_i\cap d_j|\quad 式(5)$$

其中，$|di|$表示专利 d_i 中包含的关键词数目，$|d_i\cap d_j|$表示专利 d_i 和 d_j 相同关键词数目。相似专利对中包含的相同关键词越多，则相似度越高，表明两个专利越相似；若两个专利中包含的相同关键词越少，则相

似度越低，表明这两个专利越不相似。

最终，根据相似专利对的相似度和随机专利对的相似度差值，形成相似度比（similarity difference，*SD*）：

$$SD = sim(d, d_c) - sim(d, d_r) \quad \text{式(6)}$$

表明相似专利对具有的相同关键词越多，随机专利对具有的相同关键词越少，抽取的关键词更为合理。

4 评价模型有效性实验

本部分通过锂离子电池专利数据，以验证提出的基于信息增益与相似度的评价模型的有效性。

4.1 数据来源

锂离子电池由日本 SONY 公司于 1990 年研制成功并实现商品化，锂电池以其工作电压高、能量密度高、循环寿命长、自放电低、无记忆效应、无污染、安全性能好等独特优势，从一出现就成为电化学领域的研究热点。目前已广泛应用于移动电话、便携式计算机、摄像机、照相机、电动工具等方面。随着能源危机和环境污染等问题的日益突出，开发可持续发展新能源成为当务之急，锂离子电池作为一种新型高能绿色电池备受关注。近年来随着锂离子电池技术进步，更是被广泛应用于电动汽车和储能电站等各方面。包括中国在内的主要国家已提出了明确的燃油车禁售时间表，各主流汽车厂商投入巨资来开发搭载锂电池的新能源汽车。锂离子电池关键材料主要是位于整个锂离子电池产业链中游的正极材料、负极材料、电解液、隔膜。因此，本文基于中国国家知识产权局专利数据库，检索锂离子电池中国发明专利，从检索到的发明专利中分别抽取 2 000 个专利标题与摘要，形成 C_1、C_2、C_3 和 C_4 4 个类别锂离子电池专利数据集。数据集信息如表 2 所示：

表 2 数据集信息

类别名	检索条件	类别说明	数量/个
C_1	（摘要 =‘锂离子电池’ and‘正极’）and（ IPC = H01M（用于直接转变化学能为电能的方法或装置，例如电池组））	锂离子电池正极材料	1 000
C_2	（摘要 =‘锂离子电池’ and‘负极’）and（ IPC = H01M（用于直接转变化学能为电能的方法或装置，例如电池组））	锂离子电池负极材料	1 000
C_3	（摘要 =‘锂离子电池’ and‘电解液’）and（ IPC = H01M（用于直接转变化学能为电能的方法或装置，例如电池组））	锂离子电池电解液	1 000
C_4	（摘要 =‘锂离子电池’ and‘隔膜’）and（ IPC = H01M（用于直接转变化学能为电能的方法或装置，例如电池组））	锂离子电池隔膜	1 000

4.2 数据处理

首先，对收集的专利语料进行预处理。由于中文文本的词与词之间没有间隔，为了使计算机能够识别词语，需要对中文专利文本进行分词预处理。此外，专利文本集中包含一些使用频率高但信息含量少的词，如“的”“是”等。解决该问题的方法是利用停用词表将这些词语从专利文本中剔除。最后，预处理工作还包括英文大小写格式转换、去除特殊符号等工作。

实验从 $C_1 - C_4$ 中分别选择 50 个包含引文的专利作为目标专利，并将目标专利的引用专利作为其相似专利，如表 3 所示，由 3 位领域专家为目标专利和引用专利标注 8 个关键词，使用两两交集作为最终关键词标注结果，并对人工标注结果使用 Kappa 值进行评测，Kappa 得分大于 0.8，表明标注数据的有效性。

表 3 人工标注专利关键词 （单位/个）

类别	目标专利数	引用专利数
C_1	50	50
C_2	50	50
C_3	50	50
C_4	50	50

在内部评价方法验证中，生成每个目标专利的候选关键词，计算人工标注关键词的信息增益的平均值，以及剩下的非关键词的信息增益平均值，其中式(4)中 $\alpha = 0.01$。

在外部评价方法验证中，选择目标专利与对应的引用专利，形成相似专利对，然后选择与目标专利不同类主题的其他目标专利形成随机专利对，利用专家标注关键词，计算专利相似度差。

4.3 结果分析

4.3.1 内部评价方法

图 6 为内部评价方法的实验结果。由图 6 可见，在 C_1、C_2、C_3 和 C_4 4 个类别中，关键词信息增益均值分别为 10.51、9.43、9.38 和 12.61，非关键词的信息增益均值为 4.04、4.26、2.23 和 5.11，表明关键词的信息增益显著高于非关键词的信息增益，说明专利关键词通常携带更多的信息量，能够区分不同类别间以及本类别内不同专利文献，实验结果表明利用信息增益评估专利关键词的有效性。

表 4 给出类别 C_1 锂离子电池正极材料中发明专利《碳包覆三元正极材料的制备方法及该碳包覆三元正极材料》（申请号 CN201310433513.7）的各候选关键词的信息增益值。该发明专利针对锂离子电池三元正极材料改性问题，提出一种采用有机碳包覆三元正极材料的方法。根据该专利标题和摘要生成候选关键词。

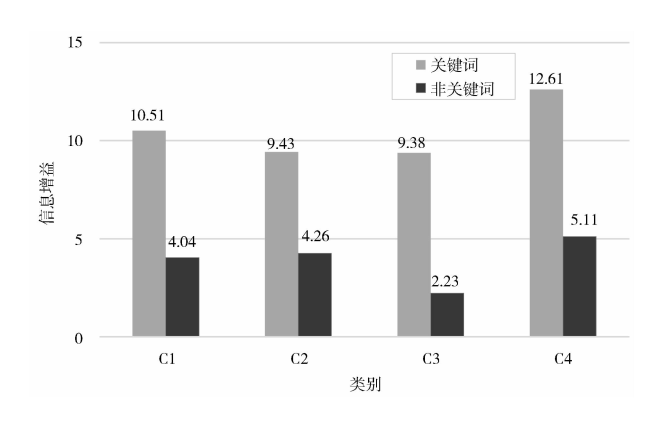


图 6　内部评价方法实验结果

表 4　内部评价方法实例

序号	候选关键词	类别信息增益	文档信息增益	信息增益
1	**镍盐**	2.32	6.94	16.19
2	**导电碳**	2.32	6.90	16.10
3	**碳包覆**	2.32	6.60	15.40
4	**前驱体**	2.32	6.43	15.01
5	**有机碳源**	2.32	4.04	9.44
6	**三元正极材料**	0.62	8.29	5.23
7	**正极材料**	0.49	8.85	4.43
8	**锂离子电池**	0.59	7.16	4.30
9	钴盐	0.43	9.41	4.14
10	锰盐	0.73	5.08	3.77
11	化合物	0.52	4.30	2.28
12	真空	0.21	5.60	1.23
13	网络状	0.28	3.73	1.08
14	通道	0.74	1.27	0.96
15	媒介	0.10	2.59	0.29
16	导电	0.07	2.74	0.22
17	倍率	0.03	4.44	0.18
18	混合物	0.12	1.16	0.15
19	性能	0.04	1.12	0.01

注：粗体表示金标准关键词

由表 4 可见，"三元正极材料""有机碳源""前驱体"等关键词具有较大的类别信息增益和较大的文档信息增益，表明关键词携带较多的信息量，能够区分不同类别，也能区分同类别的不同专利。相反，"化合物""混合物"等词因在各类别以及同类别的不同专利中均有较多的出现，因此，这些词的类别信息增益和文档信息增益值均较小，携带的信息量较少，不能体现专利的新颖性和创造性，因此具有较小的信息增益。由此可见，基于信息增益的内部评价方法能够有效地评估专利中词语携带的信息量，可作为专利关键词的判定指标。

4.3.2　外部评价方法

外部评价方法结果如图 7 所示。由图 7 可见，在 C_1-C_4 类别中，相似度差分别为 1.92、1.73、1.79 和 2.34，表明若使用正确的关键词表示专利，则相似专利对的相似度大于随机专利对的相似度。结果表明通过构建目标专利的相似专利对和随机专利对，正确的关键词能够反映专利主题，从而使得引用对比随机对具有更高的相似度。实验结果表明外部评价方法的有效性。

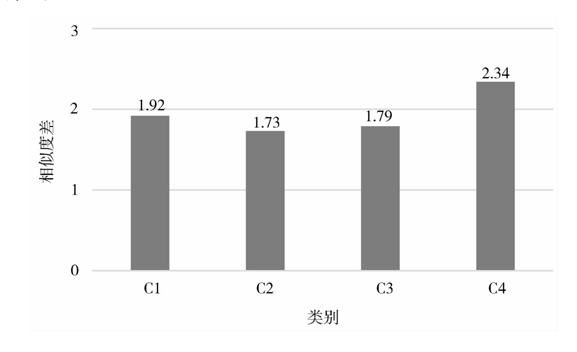


图 7　外部评价方法实验结果

表 5 为外部评价方法实例。其中目标专利为《碳包覆三元正极材料的制备方法及该碳包覆三元正极材料》（申请号 CN201310433513.7）。其相似专利为其引用专利《一种锂离子电池三元正极材料的制备方法》（申请号 CN201110314584.6），该发明专利为改进锂离子三元正极材料的不足，制备无团聚、形貌规则的单晶三元正极材料，同时对三元正极材料进行掺杂和表面包覆，使得该正极材料具有较好的循环性能和较高的安全性能，该引用专利的人工标注关键词如表 5 所示：

表 5　外部评价方法实例

专利类型	专利名	关键词	与目标专利相同关键词
目标专利	《碳包覆三元正极材料的制备方法及该碳包覆三元正极材料》	镍盐 导电碳 碳包覆 前驱体 有机碳源 三元正极材料 正极材料 锂离子电池	
引用专利	《一种锂离子电池三元正极材料的制备方法》	锂离子电池 正极材料 三元正极材料 碳链 有机添加剂 前驱体 单晶 表面包覆	正极材料 三元正极材料 锂离子电池
随机专利	《锂离子电池负极材料的制备方法、锂离子电池负极及锂离子电池》	锂离子电池 负极材料 石墨碳 等离子体 浸润性 电解液 压实密度 能量密度	锂离子电池

从表 5 可以看出,目标专利的随机专利为类别 C_2 锂离子电池负极材料中随机抽取的专利《锂离子电池负极材料的制备方法、锂离子电池负极及锂离子电池》(申请号 CN201210092946.6)。该发明专利提供了一种锂离子电池负极材料的制备方法,使得锂离子电池负极对电解液具有良好的浸润性。结果表明利用关键词表示专利时,比起随机专利,相似专利具有更多共同的专利关键词,表明使用本文提出的外部评价方法评估专利关键词的有效性。

5　评价模型应用实证

本部分应用第 3 部分提出的评价模型,比较 3 种不同专利关键词抽取策略进行评价模型应用实证研究。

具体地,专利文本包括标题、摘要、权利要求书和说明书等 4 个部分。其中,标题和摘要是对发明的概要性描述,指出发明所属领域、需要解决的技术问题、发明的主要特征和用途,文字简短,主要为专利检索提供方便途径,不具有法律效力。权利要求书是一种法律文书,描述发明的技术特征,包含体现专利新颖性和创造性的全部必不可少的技术手段或技术方法,并据此确定专利保护范围及进行专利侵权判定,是专利的核心部分。说明书是对发明的具体说明,是对权利要求书的支持,对于权利要求中的每个必要技术特征,均需要在说明书中给出详细说明,通常包括技术领域、背景技术、发明内容、附图说明、实施条件等内容。目前的专利关键词抽取研究通常集中于分析专利标题和摘要,以抽取专利关键词。本部分则尝试从专利不同部分选取候选关键词,使用经典的 TF-IDF 抽取候选关键词,使用本文提出的方法评估不同抽取策略。具体抽取策略如表 6 所示:

表 6　3 种专利关键词抽取策略

策略名	候选关键词生成部分	排序方法
abstract	标题,摘要	TF-IDF
claim	标题,摘要,权利要求书	TF-IDF
all	标题,摘要,权利要求书,说明书	TF-IDF

为此,实验以第 4.1 部分数据为基础,下载目标专利、引文专利的标题、摘要、权利要求书和说明书等文本信息,进行分词等预处理工作,使用词性匹配方法,根据不同策略,分别从标题、摘要、权利要求书和说明书中生成候选关键词,然后使用 TF-IDF 排序方法计算各个候选关键词的权重,将排名最高的八个候选关键词作为关键词。根据不同抽取策略得到的关键词,使用内部评价和外部评价方法进行评估。

内部评价方法应用结果如图 8 所示。由图 8 可见,在 3 种专利关键词抽取策略中,在 C_1、C_2、C_3 和 C_4 类别中,claim 策略的信息增益值均最大,分别为 9.42、8.55、8.61 和 10.07,表明使用 claim 策略进行关键词抽取优于 abstract 和 all 抽取策略。

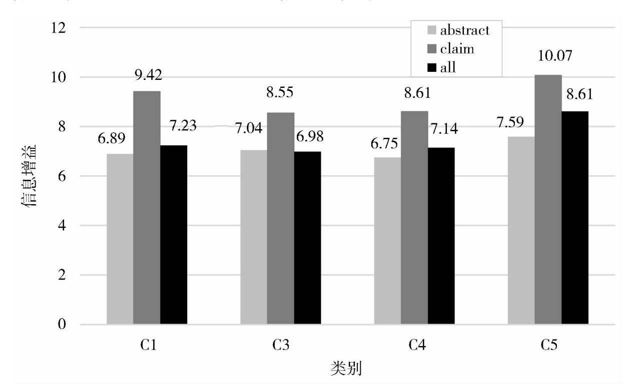


图 8　内部评价方法应用结果

表 7 列出了目标专利《碳包覆三元正极材料的制备方法及该碳包覆三元正极材料》(申请号 CN201310433513.7)使用 3 种策略抽取关键词信息增益和平均信息增益。由表 7 可见,使用 claim 策略抽取的关键词平均信息增益最高,表明使用 claim 策略抽取的关键词具有更高的准确性。

表 7　内部评价方法应用实例

关键词抽取策略	抽取的关键词	信息增益	平均信息增益
abstract	**导电碳**	16.10	6.69
	碳包覆	15.40	
	前驱体	15.01	
	三元正极材料	5.23	
	网络状	1.08	
	媒介	0.29	
	导电	0.22	
	混合物	0.15	
claim	**导电碳**	16.10	9.64
	镍盐	16.07	
	碳包覆	15.40	
	前驱体	15.01	
	三元正极材料	5.23	
	有机碳源	5.10	
	钴盐	4.14	
	分散体系	0.08	
all	**导电碳**	16.10	8.23
	碳包覆	15.40	
	前驱体	15.01	
	有机碳源	9.44	
	三元正极材料	5.23	
	正极材料	4.43	
	分散体系	0.08	
	材料	0.17	

注:粗体表示金标准关键词

图 9 为外部评价方法应用结果，在 C_1、C_2、C_3 和 C_4 类别中，使用 claim 策略的专利关键词抽取获得最高的相似度差，分别为 1.51、1.37、1.25 和 1.49，实验结果表明使用 claim 策略能够获得最好的关键词抽取结果。

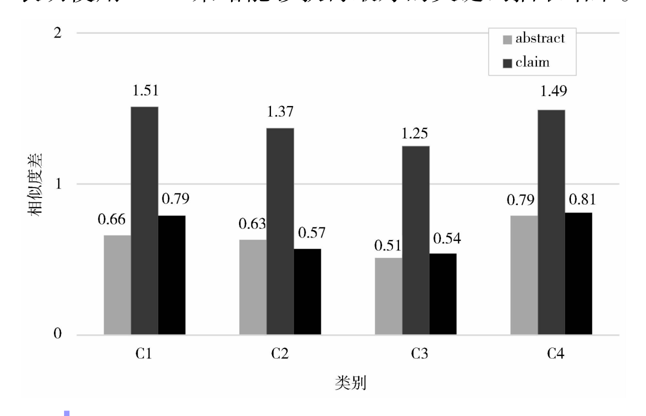


图 9　外部评价方法应用结果

表 8 为外部评价方法应用实例。目标专利为《碳包覆三元正极材料的制备方法及该碳包覆三元正极材料》(申请号 CN201310433513.7)，其引用专利为《《一种锂离子电池三元正极材料的制备方法》(申请号 CN201110314584.6)，其随机专利为《锂离子电池负极材料的制备方法、锂离子电池负极及锂离子电池》(申请号 CN201210092946.6)。使用 3 种抽取策略抽取的专利关键词以及相似度比，claim 抽取策略具有最大的相似度差，表明 claim 抽取的关键词最好地刻画了专利主题内容。

表 8　外部评价方法应用实例

抽取策略	目标专利关键词	引用专利关键词	随机专利关键词	相似度差
abstract	导电碳 碳包覆 前驱体 三元正极材料 网络状 媒介 导电 混合物	三元正极材料 碳链有机添加剂 原料偏析 高温反应过程 胶体磨 研磨过程 混料均匀度研磨	锂离子电池 石墨碳 填充量 负极材料 浸润性 浸润程度 等离子体处理装置	1－0 ＝ 1
claim	导电碳 镍盐 碳包覆 前驱体 三元正极材料 有机 碳源 钴盐 正极材料	三元素中间体 三元正极材料 镍钴锰 可溶性盐 碳链有机添加剂 胶体磨 正极材料 锂源	等离子体 锂离子电池 处理装置 石墨碳 导电基体 硫化氢 负极材料 氨气	2－0 ＝ 2
all	导电碳 碳包覆 前驱体 有机碳源 三元正极材料 正极材料 分散体系 材料	气流磨 胶体磨 保温 三元正极材料 摩尔 产物 成型物料 压片机	石墨碳 样品 等离子体 处理装置 锂离子电池 导电基体 氨气 浸润性	1－0 ＝ 1

综上所述，从内部评价方法和外部评价方法应用结果可见，专利文献不同于其他文献，仅仅利用标题和摘要抽取关键词，可能遗漏一些关键词。究其原因，一些专利摘要存在书写过于简单的问题，且专利摘要过于简短，无法使用 TF-IDF 方法很好地抽取专利关键词。专利说明书内容详实，包含较多的实例和具体步骤和技术细节，但同时也包含较多的噪声数据，从而导致抽取的专利关键词准确性不尽理想。相反，专利权利要求书既是技术文书，也是法律文书，体现了专利的新颖性和创造性，是专利的核心部分。因此，不同于其他文献，实验表明，从权利要求书中抽取专利关键词比从摘要和说明书中抽取关键词具有更好的准确性。

6　结论

专利关键词是表明专利文献主题内容的一组词或者短语，被广泛应用于专利分类、新兴技术监测、专利检索、专利聚类等专利分析之中。利用计算机自动、高效、准确地抽取专利关键词是一个重要的研究课题。目前的专利关键词抽取研究主要集中于专利关键词抽取算法的改进，而各种改进算法有效性评价通常采用算法抽取的关键词与专家人工标注的关键词进行匹配。然而，依靠专家人工标注专利关键词费时费力且标注数量有限，存在领域局限性，语言依赖性、主观性等问题，使得专利关键词抽取算法无法有效地被评价，阻碍专利关键词抽取算法的进一步深入研究。

本文从内部评价和外部评价两个角度，提出基于信息熵的内部评价模型与基于相似度的外部评价模型，以缓解目前专利关键词抽取算法评价方法的不足。基于信息熵的内部评价方法分别考察每个被抽取的关键词，使用类别信息增益和文档信息增益度量每个被抽取的关键词所携带的信息量，从而刻画算法抽取关键词的创造性和新颖性。基于相似度的外部评价方法则用算法抽取的关键词集表示专利，比较相似专利对相似度与随机专利对相似度，评价算法抽取的关键词刻画专利主题内容的准确性。实验结果验证了本文提出的基于信息增益与相似度的评价模型的有效性。进一步地，利用本文提出的评价模型，进行实际应用，评估 3 种专利关键词抽取策略，实验表明本文提出评价模型具有有效性与可行性。该评价模型主要用于评价专利关键词抽取算法的有效性，但模型也可用于学术文献等关键词抽取算法的评价，具有一定的适用性。

然而，本文提出的评价模型也存在需要进一步改进之处。首先，基于信息增益的内部评价模型主要评

估了抽取的关键词刻画专利的新颖性和创造性的能力，但是，在专利中，存在一些数值上的创新，如电池存储容量提高了 20% 等，这需要进一步完善评价模型；其次，本文提出的评价模型需要构建相应的专利集，而构建专利集对评价模型稳定性的影响也需要进一步研究；最后，本文提出的评价模型对于学术文献等关键词抽取算法的适用性也需要在将来的实验中加以验证。

参考文献：

[1] HU J, LI S, YAO Y, et al. Patent keyword extraction algorithm based on distributed representation for patent classification[J]. Entropy, 2018, 20(2): 104-124.

[2] JOUNG J, KIM K. Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data [J]. Technological forecasting and social change, 2017, 114: 281-292.

[3] 周胜生. 关键词在专利文献检索中的应用[J]. 情报理论与实践, 2010, 33(5):67-70.

[4] 王坤, 王京安, 汤月, 等. 基于专利和科技论文的技术机会识别研究——以金属 3D 打印技术为例[J]. 科技管理研究, 2018(7):73-79.

[5] FIROOZEH N, NAZARENKO A, ALIZON F, et al. Keyword extraction: Issues and methods[J]. Natural language engineering, 2020, 26(3): 259-291.

[6] RISTAD E S, YIANILOS P N. Learning string-edit distance[J]. IEEE transactions on pattern analysis and machine intelligence, 1998, 20(5): 522-532.

[7] DAGAN I, PEREIRA F. Similarity-based estimation of word cooccurrence probabilities[EB/OL]. [2021-10-28]. https://arxiv.org/pdf/cmp-lg/9405001.pdf.

[8] 章成志, 周冬敏. 自动标引通用评价模型研究[J]. 情报学报, 2009, 28(1): 40-47.

[9] 俞琰, 尚明杰, 赵乃瑄. 权利要求特征驱动的专利关键词抽取方法[J]. 情报学报, 2021, 40(6): 610-620.

[10] 马慧芳, 刘芳, 夏琴, 等. 基于加权超图随机游走的文献关键词提取算法[J]. 电子学报, 2018, 46(6): 1410-1414.

[11] 王志宏, 过弋. 基于词句重要性的中文专利关键词自动抽取研究[J]. 情报理论与实践, 2018, 41(9): 123-129.

[12] SINGHAL A, KASTURI R, SRIVASTAVA J, et al. Leveraging web resources for keyword assignment to short text documents[EB/OL]. [2021-10-28]. https://arxiv.org/ftp/arxiv/papers/1706/1706.05985.pdf.

[13] VOORHEES E M. The TREC-8 question answering track report [EB/OL]. [2021-10-28]. https://citeseerx.ist.psu.edu/viewdoc/download? doi=10.1.1.38.6392&rep=rep1&type=pdf.

[14] FLORESCU C, CARAGEA C. Positionrank: an unsupervised approach to keyphrase extraction from scholarly documents[C]//Proceedings of the 55th annual meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 1105-1115.

[15] ZHANG Y, CHANG Y, LIU X, et al. Mike: keyphrase extraction by integrating multidimensional information[C]//Proceedings of the 2017 ACM on conference on information and knowledge management. New York: ACM, 2017: 1349-1358.

[16] BUCKLEY C, VOORHEES E M. Retrieval evaluation with incomplete information[C]//Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2004: 25-32.

[17] LIU Z, HUANG W, ZHENG Y, et al. Automatic keyphrase extraction via topic decomposition[C]//Proceedings of the 2010 conference on empirical methods in natural language processing. Stroudsburg: ACL, 2010: 366-376.

[18] ZHANG K, XU H, TANG J, et al. Keyword extraction using support vector machine[C]//Proceedings of the 6th international conference on advances in Web-age information management conference. Berlin: Springer-Verlag, 2006: 85-96.

[19] TURNEY P D. Mining the Web for lexical knowledge to improve keyphrase extraction: learning from labeled and unlabeled data [EB/OL]. [2021-10-31]. https://arxiv.org/ftp/cs/papers/0212/0212011.pdf.

[20] KIM S N, MEDELYAN O, KAN M-Y, et al. SemEval-2010 task 5: automatic keyphrase extraction from scientific articles[EB/OL]. [2021-10-31]. https://aclanthology.org/S10-1004.pdf.

[21] AUGENSTEIN I, DAS M, RIEDEL S, et al. SemEval 2017 task 10: ScienceIE-Extracting keyphrases and relations from scientific publications[EB/OL]. [2021-10-31]. https://arxiv.org/pdf/1704.02853.pdf.

[22] RODRIGUEZ A, KIM B, TURKOZ M, et al. New multi-stage similarity measure for calculation of pairwise patent similarity in a patent citation network[J]. Scientometrics, 2015, 103(2):565-581.

[23] 李睿, 张玲玲, 郭世月. 专利同被引聚类与专利引用耦合聚类的对比分析[J]. 图书情报工作, 2012, 56(8):91-95.

[24] LU Y, XIONG X, ZHANG W, et al. Research on classification and similarity of patent citation based on deep learning[J]. Scientometrics, 2020, 123(1): 813-839.

作者贡献说明：

俞琰：提出研究思路，设计研究方案，进行实验，撰写论文；
鞠鹏：分析数据，修改论文；
尚杰明：收集数据，清洗数据。

Research on the Evaluation Method of Patent Keyword Extraction Algorithm Based on Information Gain and Similarity

Yu Yan[1,2] Ju Peng[1] Shang Mingjie[1]
[1] Institute of the Information Management and Technology, Nanjing Technology University, Nanjing 210009
[2] School of Electronics and Computer, Chengxian College, Southeast University, Nanjing 211816

Abstract: [**Purpose/significance**] Aiming at the problems existing in the evaluation of patent keyword extraction algorithm, which mainly uses the extracted keywords to match the keywords manually labeled by experts, an evaluation model of patent keyword extraction algorithm based on information gain and similarity is proposed. [**Method/process**] The proposed evaluation model evaluated the accuracy of the patent keyword extraction algorithm from intrinsic and extrinsic levels. The intrinsic evaluation model measured the information gain of each keyword extracted by the evaluation algorithm to evaluate the novelty and creativity of the extracted keywords. The extrinsic evaluation model used the keyword set extracted by the evaluation algorithm to represent the patents, and measured the effectiveness of the keywords extracted by the algorithm to describe the patent topic by calculating the similarity of relevant patents. [**Result/conclusion**] Through the validation experiment of the evaluation model and the empirical research on the application of the evaluation model, the results show that the evaluation model based on information gain and similarity is feasible and effective.

Keywords: patent　keyword extraction　evaluation　information gain　similarity